Indonesian Journal of Electrical Engineering and Computer Science
Vol. 44, No. 1, October 2026, pp. 164~176
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v44.i1.pp164-176

Improving Naïve Bayes sentiment classification with SMOTE–Tomek hybrid resampling

Yo Ceng Giap, Christoper Andrian
Faculty of Science and Technology, Buddhi Dharma University, Tangerang, Indonesia

Article Info

Article history:

Received Sep 12, 2025
Revised Jan 8, 2026
Accepted Sep 10, 2026

Keywords:

Classification
Naïve Bayes
Ruangguru
Sentiment analysis
SMOTE
Tomek Links

ABSTRACT

Class imbalance remains a critical challenge in sentiment classification, particularly in reviews of digital education applications where minority sentiments are underrepresented. This study investigates a hybrid resampling framework that integrates Naïve Bayes with synthetic minority oversampling technique (SMOTE)–Tomek Links to improve sentiment classification under imbalanced conditions. Experiments were conducted on 9,700 Ruangguru application reviews collected from Google Play and represented using term frequency–inverse document frequency (TF-IDF) features. Results show that the baseline Naïve Bayes model achieves 83.52% accuracy but performs poorly on minority classes, while the proposed hybrid approach improves accuracy to 85.33% and yields more balanced macro- and weighted-average F1-scores. Confidence interval analysis (95%) further indicates improved performance stability compared to the baseline model. Given its low computational complexity and preprocessing-only resampling strategy, the proposed framework provides an efficient and scalable solution for reliable sentiment monitoring in digital education platforms.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Yo Ceng Giap
Faculty of Science and Technology, Buddhi Dharma University
Tangerang City, Banten Province, Indonesia
Email: cenggiap@ubd.ac.id

1. INTRODUCTION

The rapid growth of digital platforms has transformed how users express opinions and experiences online. In the education sector, mobile learning applications such as Ruangguru have become essential services for millions of students in Indonesia. Reviews published on the Google Play Store not only reflect user satisfaction but also provide valuable data for evaluating service quality and supporting data-driven decision-making in digital education environments [1], [2]. Consequently, sentiment analysis of user-generated reviews has become increasingly important for extracting meaningful insights within the educational technology ecosystem.

Sentiment analysis, an essential function in natural language processing (NLP), categorizes opinions as positive, neutral, or negative [3], [4]. Nonetheless, its efficacy is frequently limited by class imbalance, in which positive ratings predominate while neutral and negative opinions are underrepresented [5]. This disparity biases classifiers, such as Naïve Bayes, toward the majority class, leading to inadequate identification of the minority class. Prior research indicates that Naïve Bayes achieves only 53%–69% accuracy on imbalanced hotel review datasets, due to its inability to effectively identify neutral and negative attitudes [6]. Comparable constraints have been noted in software requirement classification, characterized by disproportionate distributions of functional and non-functional categories [7], underscoring the algorithm's susceptibility to biased input [8]. While Naïve Bayes is computationally economical and performs well in text categorization, its efficacy diminishes on imbalanced sentiment datasets [9], [10].

In educational application reviews, the imbalance problem is further intensified by linguistic characteristics, including informal expressions, code-mixing between Indonesian and English, domain-specific educational terminology, and contextual nuances such as negation. These features distinguish educational reviews from typical e-commerce or hospitality datasets, making reliable sentiment classification more difficult.

To address class imbalance, resampling techniques such as the synthetic minority oversampling technique (SMOTE) are commonly used to increase minority-class representation [11]. Nonetheless, SMOTE may generate overlapping samples and noise, thereby undermining class boundaries [12]. Tomek Links addresses this issue by eliminating confusing majority-class samples near decision boundaries, thereby enhancing data integrity [13]. The hybrid SMOTE–Tomek Links method combines oversampling and data cleansing to create a more balanced dataset, thereby improving classification performance across many domains [14].

Notwithstanding these advancements, research that explicitly combines Naïve Bayes with SMOTE–Tomek Links for sentiment analysis in educational contexts is limited. The majority of current research focuses on e-commerce, hospitality, numerical forecasting, or fraud detection rather than on educational platforms with unique linguistic settings. Although recent studies report performance gains using this hybrid approach in domains such as air quality prediction and fraud detection [15], [16], its effectiveness in educational sentiment analysis has not been sufficiently explored.

Therefore, this study investigates whether a hybrid resampling strategy using SMOTE–Tomek Links can improve minority-class sentiment recognition in educational application reviews without substantially degrading majority-class performance when combined with a lightweight Naïve Bayes classifier. This study presents a Naïve Bayes-based sentiment classification framework with hybrid resampling, applied to Ruangguru reviews as a sample digital education platform in Indonesia, and assesses its performance through accuracy, precision, recall, and F1-score metrics. This work's contribution is characterized as domain-specific and focused on efficiency, rather than asserting wide generality.

2. RELATED WORK

Class imbalance has long been recognized as a critical issue in text classification and sentiment analysis. Naïve Bayes is commonly employed due to its simplicity and effectiveness; however, its assumption of conditional independence and its susceptibility to class imbalance often result in diminished performance on imbalanced datasets [17]. This phenomenon frequently occurs in real-world textual data, where positive emotion predominates over neutral and negative categories [18]. To mitigate this problem, various resampling techniques have been proposed, including oversampling methods such as SMOTE and undersampling methods such as Tomek Links, as well as their hybrid combinations [19], [20].

Empirical research repeatedly demonstrates that resampling enhances the identification of the minority class. Putra *et al.* [6] found that SMOTE improves the identification of neutral and negative sentiments in hotel reviews, although individual classifiers such as Naïve Bayes and SVMs exhibit bias towards the majority class. Arsanto *et al.* [21] documented a significant improvement in minority-class performance by integrating Naïve Bayes with SMOTE–Tomek Links for credit card fraud detection; however, their research was confined to the banking sector. Suandi *et al.* [22] demonstrated that integrating SMOTE, ADASYN, and Tomek Links increased sentiment analysis accuracy to 89% without relying on educational datasets or Naïve Bayes-based hybrid resampling.

Hybrid resampling has demonstrated efficacy in non-textual domains as well. Chandra *et al.* [16] demonstrated that SMOTE–Tomek Links reduced noise and improved minority-class accuracy in air quality prediction relative to SMOTE alone, whilst Taskiran *et al.* [14] emphasized its efficacy in alleviating class overlap, particularly in environmental data [14]. Methodological reviews further emphasize that SMOTE often generates noisy or redundant samples and should be combined with data-cleaning techniques such as Tomek Links [12]. Extensions such as Select-SMOTE with LightGBM [23], SMOTE–Tomek in software requirement classification [7], and fault detection using Naïve Bayes [15] confirm the general effectiveness of hybrid resampling, although these studies remain outside the scope of educational sentiment analysis.

The explicit integration of Naïve Bayes with SMOTE–Tomek Links in sentiment analysis applications is scarce. Current studies encompass sentiment analysis of the by.U application utilizing Naïve Bayes with SMOTE, which indicated a slight accuracy enhancement from 84.42% to 85.83% in a non-educational setting [24], and the implementation of SMOTE–Tomek Links with SVM in Kitabisa reviews, attaining elevated accuracy and F1-score [25]. In the educational sector, Naïve Bayes combined with SMOTE has been used to assess sentiment on e-sports in the curriculum, achieving 70.32% accuracy and an AUC of 0.954, but without hybrid resampling [26].

Previous research demonstrates that hybrid resampling techniques, such as SMOTE–Tomek Links, enhance classification efficacy across multiple domains. Nonetheless, the systematic assessment of Naïve

Bayes integrated with SMOTE–Tomek Links for sentiment analysis in educational application reviews has yet to be thoroughly investigated. This study addresses this gap by focusing on educational sentiment data and assessing classification balance using comprehensive performance metrics.

3. METHOD

Figure 1 shows how the proposed SMOTE-Tomek Links hybrid method was tested in a systematic way to see if it could improve Naïve Bayes's performance in sentiment classification. It describes the experimental setup, how the data were collected, the steps taken to clean the data, and the evaluation metrics used to assess the proposed method's performance. The main goal is to show how combining SMOTE to oversample minority classes with Tomek Links to clean noisy, overlapping data can make imbalanced datasets much easier to work with and improve the performance of the Naïve Bayes classifier.

This meticulous methodology ensures that the insights obtained are substantial and can be effectively applied to practical sentiment analysis challenges, particularly in domains such as product evaluations and social media, where sentiment distributions often exhibit irregularities [27]. This research utilizes a multi-class methodology, categorizing texts as positive, negative, or neutral, which is essential for thorough sentiment analysis [28].

To accomplish this, the subsequent research methods meticulously tackle data acquisition, preprocessing, feature extraction, model training, resampling, and evaluation matrices to establish a resilient framework for performance assessment.

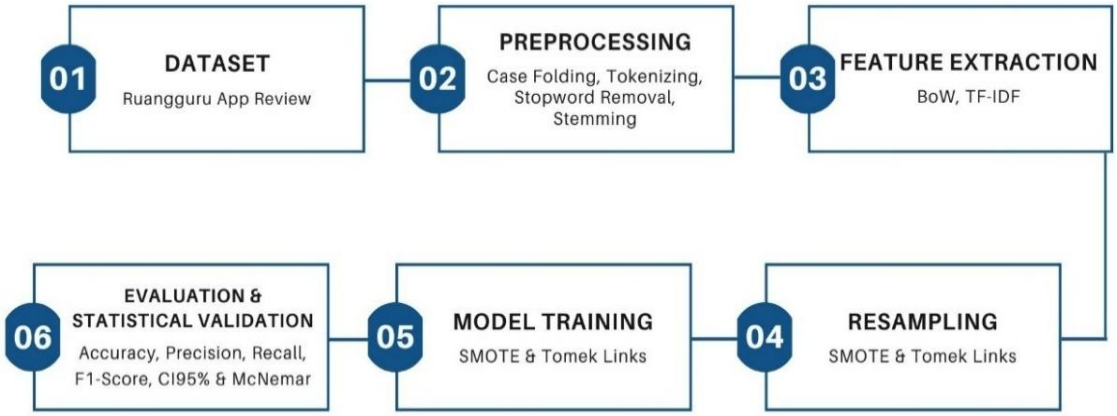

Figure 1. Research methods

A. Dataset

This study uses a dataset consisting of user reviews of the Ruangguru app obtained from the Google Play Store. A total of 9,700 data points were collected. The reviews were labeled into three sentiment categories: positive, neutral, and negative. The data distribution was unbalanced: the positive class (>60%), the neutral class (around 25%), and the negative class (only 15%). This imbalance was the main challenge in building a sentiment analysis model.

B. Preprocessing

The preprocessing process is carried out to prepare text data before modeling, including:

- Case Folding: change all letters to lowercase.
- Tokenizing: breaking review sentences into words (tokens).
- Stopword Removal: removing common words such as “and”, “or”, “that” have no significant meaning.
- Stemming: return the word to its basic form, for example, “belajarannya” → “belajar”.

C. Feature Extraction

Text representation is done by:

- Bag-of-words (BoW): counting word frequency.
- Term frequency–inverse document frequency (TF-IDF): giving greater weight to words that are rare but important in the text. In this study, TF-IDF was used as the primary method because it is more representative in review analysis.

D. Resampling: SMOTE and Tomek Links

To address the class imbalance problem in sentiment classification, this study employs a hybrid resampling strategy that combines Synthetic Minority Oversampling Technique (SMOTE) and Tomek Links. The objective of this approach is to simultaneously improve minority class representation and reduce noisy or overlapping samples, thereby enhancing the overall classification performance of the Naïve Bayes model.

- SMOTE

  SMOTE is an oversampling technique that generates synthetic samples for minority classes by interpolating between existing minority instances and their nearest neighbors. This procedure enhances the diversity of samples from the minority class and mitigates the classifier's bias towards the majority class. This work employs SMOTE to equilibrate the distribution of sentiment categories (positive, neutral, and negative) during the training phase. This helps the classifier develop a deeper understanding of the minority class's attributes.

- Tomek links

  Tomek Links is a method for cleaning up and undersampling data that finds pairs of samples from different classes that are closest to each other. In each Tomek Link pair, the instance from the majority class is removed. This is due to the proximity of these pairs to class boundaries, which might complicate classification. Tomek Links mitigates data noise and enhances class separation by eliminating borderline samples.

- Parameter configuration of SMOTE–Tomek links

  The hybrid resampling process is done in steps. First, SMOTE is used to balance class distributions, and then Tomek Links is used to improve data quality. This order is very important because using Tomek Links before oversampling would not fix the problem of insufficient minority-class samples.

  In SMOTE, the number of nearest neighbors (k-neighbors) is established at k = 5. This is a prevalent metric in the literature on unbalanced learning, effectively balancing the generation of diverse synthetic samples while minimizing excessive class overlap. The oversampling approach guarantees uniform sample sizes among all sentiment categories. This ensures that model training is fair.

  After oversampling, Tomek Links is used to clean the data by removing the majority-class samples that form Tomek pairs. This method does not require any additional threshold parameters, as it uses only nearest-neighbor relationships to identify unclear samples. The model can benefit from better representation of the minority class and reduced noise and overlapping areas arising from oversampling when SMOTE and Tomek Links are used together.

  This study does not perform a comprehensive sensitivity analysis of alternative SMOTE parameters, including varying k-neighbor values or sampling ratios, to mitigate computational complexity. Nonetheless, investigating these parameter variations is recognized as a prospective avenue for future research.

E. Model Training

Naïve Bayes is the baseline model used. It is a simple, efficient, and effective algorithm for text classification and sentiment analysis. This model is the starting point for the resampling method. Tomek Links and SMOTE are used to improve the model.

F. Evaluation Metrics and Statistical Validation

Model performance is evaluated using standard classification metrics, namely:

- Accuracy: ratio of correct predictions to total data.
- Precision: true positive prediction rate.
- Recall: proportion of positive data correctly detected.
- F1-Score: the harmonic mean between precision and recall, critical for evaluating minority classes.

In addition to standard evaluation metrics, a 95% confidence interval (CI) analysis was used to assess the results statistically. McNemar's test was considered for paired comparisons; however, it was not used because the evaluation set was constructed differently after resampling.

4. RESULT AND DISCUSSION

This section presents the results of experiments evaluating the performance of the Naïve Bayes method for sentiment classification, both as a baseline and after applying imbalanced data handling techniques. The comparison covers four main scenarios, namely (A) Naïve Bayes without resampling, (B) Naïve Bayes with SMOTE, (C) Naïve Bayes with Tomek links, and (D) a combination of Naïve Bayes with SMOTE-Tomek Links. The analysis described not only compares accuracy, precision, recall, and F1-score

values, but also emphasizes the effectiveness of each approach in improving the detection of minority classes in sentiment data.

4.1. Naïve Bayes (baseline)

Figure 2 illustrates the distribution of sentiment classes in the original dataset. In the baseline experiment, the Naïve Bayes model was run directly on the original data without resampling. The data distribution in the figure below shows a significant class imbalance, with approximately 6,726 positive data points, 2,214 neutral data points, and only 760 negative data points.

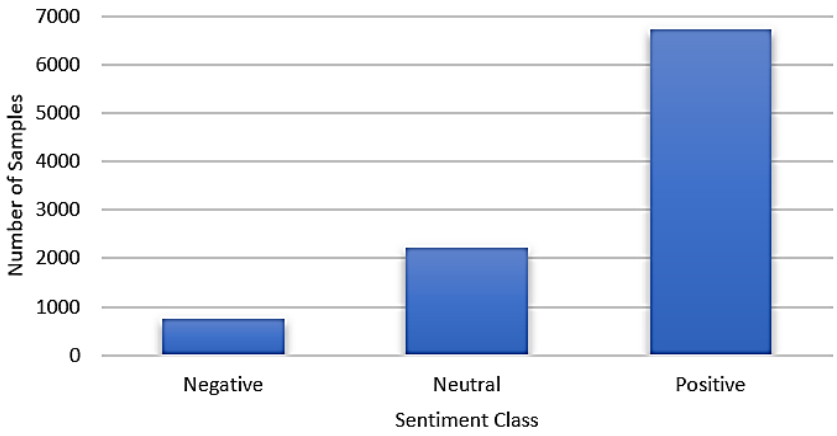

Figure 2. Amount of data

The confusion matrix results for the Naïve Bayes model (baseline) in Table 1 show that the model's performance remains significantly affected by data distribution imbalance. Out of 76 data points in the negative class, only 22 were correctly predicted, while 11 were misclassified as neutral and 43 were misclassified as positive. This shows that the model struggles to recognize the negative class, which has very little data. In the neutral class, out of 222 data points, 170 were correctly predicted, but 3 were incorrectly classified as negative and 49 as positive. Meanwhile, in the positive class, out of 673 data points, 619 were correctly predicted, while 21 were incorrectly recognized as negative and 33 as neutral. These findings show that the model has a strong tendency to acknowledge the positive class as the majority, but is weak in classifying minority classes, especially negative ones.

Table 1. Confusion matrix Naïve Bayes (baseline)

Class	Prediction results			Total
	Negative	Neutral	Positive	
Negative	22	11	43	76
Neutral	3	170	49	222
Positive	21	33	619	673

The first tests with Naïve Bayes without resampling showed that the model was still very affected by data imbalance, as shown in Table 2. The model performed well on the majority class, which was positive. It achieved a precision of 87.06%, a recall of 91.98%, and an F1-score of 89.45%. This makes sense because there was much more positive data than from the other classes. On the other hand, the negative class did very poorly, with a recall of only 28.95% and an F1-score of 36.06%. This means that most of the negative data was not correctly identified. The neutral class had an F1 score of 75.31%, which was still lower than the positive class. The overall accuracy was 83.52%, which seems pretty good, but the macro-average F1-score was only 67.83%, much lower than the weighted average of 82.64%. This difference shows that the model performs better on the majority class than on the minority class. This baseline shows that Naïve Bayes struggles with unbalanced data. To improve classification across all classes, resampling methods such as SMOTE, Tomek Links, or a combination of both are needed.

Table 2. Results of confusion matrix Naïve Bayes (baseline)

Class	Precision	Recall	F1-score
Negative	47,83%	28,95%	36,06%
Neutral	79,44%	76,58%	75,31%
Positive	87,06%	91,98%	89,45%
Accuracy			83,52%
Macro avg	71,44%	65,84%	67,83%
Weighted avg	82,24%	83,52%	82,64%

4.2. Naïve Bayes + SMOTE

Figure 3: The SMOTE algorithm addresses data imbalance in baseline experiments. As shown in the graph, before SMOTE, the data distribution was highly imbalanced, with the positive class comprising more than 6,726 data points. In comparison, the negative class had approximately 760 data points, while the neutral class had approximately 2,214. After applying SMOTE, the data in each class was balanced to a relatively similar range (6,726 data points per class).

With a more balanced data distribution, Naïve Bayes is no longer overly biased towards the majority class. The model has a greater opportunity to learn from minority-class representations, particularly negative and neutral ones. This is expected to improve recall and F1 scores for both classes, which were previously very low in the baseline. However, it should be noted that oversampling can also introduce noise, as synthetic data does not fully reflect real data.

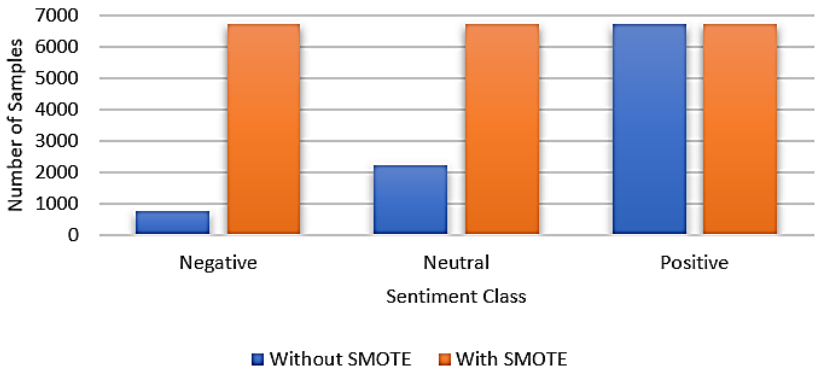

Figure 3. Comparison of the amount before and after using Naïve Bayes + SMOTE

Table 3 presents the confusion matrix results, which indicate that applying SMOTE successfully improved Naïve Bayes' performance on minority classes, particularly the negative class. Out of 672 negative data points, 657 were accurately predicted, whereas 15 were misclassified. This signifies a considerable enhancement relative to the baseline, when the recall for the negative class was previously exceedingly low. The model exhibited outstanding performance in the neutral class, accurately categorizing 479 of 673 data points. Nevertheless, it committed multiple errors, misclassifying 110 neutral data points as negative and 84 as positive. Meanwhile, in the positive class, 541 data points were correctly predicted out of 673, while 132 were incorrectly predicted as negative or neutral. Although performance in the majority class declined slightly, the improvement in the minority class made the classification results more fair and representative for all sentiment categories.

Table 3. Confusion matrix using Naïve Bayes + SMOTE

Class	Prediction results			Total
	Negative	Neutral	Positive	
Negative	657	4	11	672
Neutral	110	479	84	673
Positive	75	57	541	673

Based on the evaluation in Table 4, applying SMOTE provides a significant improvement in performance for the negative class. Precision increased to 78.03%, recall reached 97.77%, and the F1-score

reached 86.80%, which is much better than the baseline, which previously produced only 28.95% recall and 36.06% F1-score. This is consistent with the confusion matrix results, which show that most negative data points were correctly classified (657 out of 672).

In the neutral class, the results show high precision (88.70%) but lower recall (71.17%), resulting in an F1-score of 78.97%. This aligns with the confusion matrix, which shows significant prediction errors, especially when neutral data is misclassified as negative or positive. Meanwhile, in the positive class, the model's performance declined slightly compared to the baseline. Precision was 85.06%, recall was 80.39%, and the F1-score was 82.66%, indicating that although still quite good, the model was no longer as dominant in the majority class as before.

The model achieved an overall accuracy of 83.10%, a macro-average F1-score of 82.81%, and a weighted-average F1-score of 82.80%. The data indicate that class performance is now more equitable than previously, when the macro-average significantly lagged behind the weighted-average. SMOTE mitigated data imbalance by equilibrating the sample sizes across classes. This indicated that the Naïve Bayes model was no longer overly biased toward the majority class. The confusion matrix results indicate that predicting negative outcomes, which were previously difficult to identify, is now considerably easier. Even though the positive class's performance dropped slightly, the model's overall performance became more balanced between classes, as shown by a large increase in the macro-average. So, SMOTE has been shown to improve Naïve Bayes' sentiment classification on unbalanced data.

Table 4. Results of confusion matrix using Naïve Bayes + SMOTE

Class	Precision	Recall	F1-score
Negative	78,03%	97,77%	86,80%
Neutral	88,70%	71,17%	78,97%
Positive	85,06 %	80,39%	82,66%
Accuracy			83,10%
Macro avg	83,93 %	83,11%	82,81%
Weighted avg	83,56%	83,10%	82,80%

4.3. Naïve Bayes + Tomek Links

Tomek Links is applied to reduce data that has the potential to become borderline noise, which is data that lies on the boundary between classes and is prone to misclassification. Figure 4 shows that after this process, the amount of data in the negative, neutral, and positive classes decreased slightly compared to before Tomek Links was applied. The amount of negative data decreased slightly (remaining relatively small), neutral data decreased from around 2,214 to around 2,062, while positive data decreased from more than 6,726 to around 6,560.

With this reduction in data, the Naïve Bayes model is expected to focus more on "clean," representative data patterns, thereby reducing the risk of misclassification caused by ambiguous data. However, because this method only cleans the data without adding data to the minority class, the data distribution remains unbalanced. This means that the positive class remains dominant, while the negative class remains a minority class with far less representation.

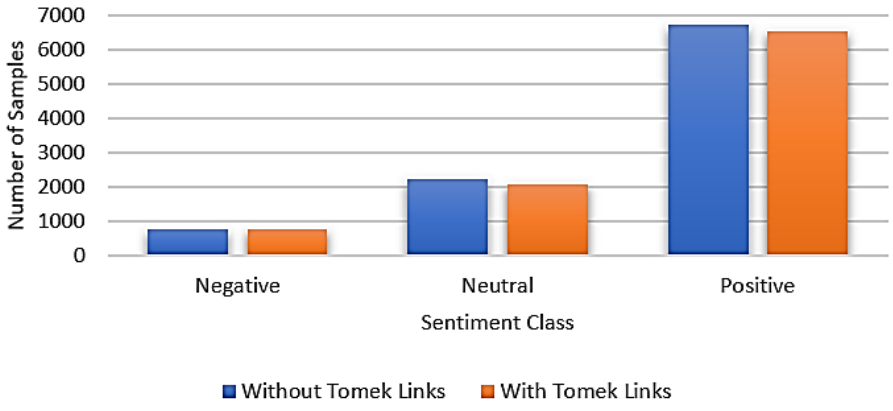

Figure 4. Comparison of the amount before and after using Naïve Bayes + Tomek Links

Table 5 shows the confusion matrix results, which indicate that applying Tomek Links to Naïve Bayes has not improved the model's performance on minority classes. In the negative class, out of 76 data points, only 2 were correctly predicted, while the other 74 were misclassified as positive. A similar situation occurred in the neutral class: of 206 data points, only 39 were correctly classified, while the remaining 167 were incorrectly identified as positive. In contrast, performance in the positive class was very strong, with 656 out of 657 data points correctly predicted and only 1 incorrectly. These findings show that although Tomek Links can reduce ambiguity at class boundaries, this method does not address data imbalance. The model remains biased towards the majority class, resulting in very low performance in the minority class. Thus, Tomek Links is better suited as a complementary method, for example, in combination with SMOTE, to achieve more balanced classification across all classes.

Table 6 shows very uneven test results between classes. In the negative class, although precision reached 100%, recall was only 3%, resulting in an F1-score of 5%. This means that, among the few negative data points correctly detected, all were correctly predicted. Still, most negative data points were not recognized and were incorrectly categorized as positive. A similar situation occurred in the neutral class, with 97% precision but only 19% recall, yielding an F1 Score of 32%. This indicates that most neutral data was also misclassified, mainly as positive. Conversely, in the positive class, recall is very high (100%), and the F1-score is 84%, indicating that the model tends to classify the majority of data as positive.

Table 5. Confusion matrix using Naïve Bayes + Tomek Links

Class	Prediction results			Total
	Negative	Neutral	Positive	
Negative	2	0	74	76
Neutral	0	39	167	206
Positive	0	1	656	657

Table 6. Results of confusion matrix using Naïve Bayes + Tomek Links

Class	Precision	Recall	F1-score
Negative	100%	3%	5%
Neutral	97%	19%	32%
Positive	73%	100%	84%
Accuracy			74%
Macro avg	90%	40%	40%
Weighted avg	81%	74%	66%

Overall, the model's accuracy reached 74%, but this value does not reflect the distribution of performance across classes. This is evident in the macro-average F1-score, which is only 40%, well below the weighted average (66%). This difference confirms that the model remains highly biased towards the positive class, with a notably weak ability to detect minority classes (negative and neutral). Given the amount of data, the confusion matrix, and the test results, it can be concluded that applying Tomek Links does not improve Naïve Bayes performance on highly imbalanced data. The amount of negative and neutral data remains far lower than that of positive data, so even though ambiguous data is reduced, the model still predominates the majority class. The confusion matrix shows a dominance of positive predictions across almost all classes, and the evaluation metrics show very low recall for the minority class. Thus, Tomek Links serves only as a data-cleaning method and does not address the problem of distribution imbalance. This reinforces previous findings that this method should be used as a complement to oversampling (e.g., SMOTE-Tomek Links) to produce more balanced and fair classification across all classes.

4.4. Naïve Bayes + SMOTE + Tomek Links

The combination of the SMOTE and Tomek Links methods addresses two problems simultaneously: data distribution imbalance and the presence of borderline noise data. Figure 5 shows that, before this method was applied, the data distribution was highly imbalanced: the positive class was dominant, with more than 6,726 data points, while the neutral class had around 2,214. The negative class only had around 760 data points. After combining SMOTE and Tomek Links, the number of data in each class was balanced to around 6,697 for neutral and positive, and 6,726 for negative.

Under balanced data conditions, Naïve Bayes provides a more proportional representation of each class, thereby increasing the model's ability to recognize patterns in minority classes. SMOTE adds synthetic data to minority classes, while Tomek Links helps remove ambiguous data that could lead to misclassification. The combination of the two makes the distribution fairer while improving the quality of the training data.

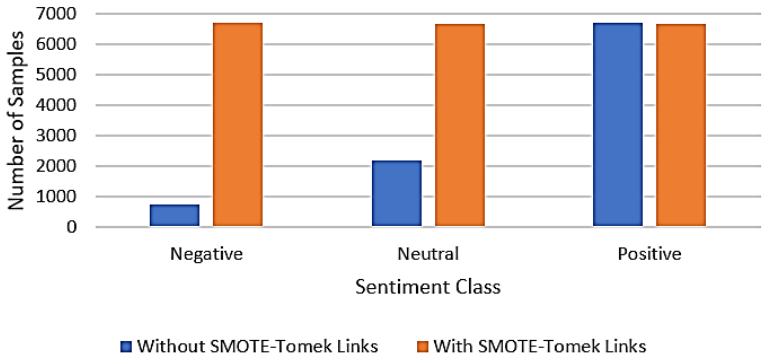

Figure 5. Comparison of the amount before and after using Naïve Bayes + SMOTE-Tomek Links

Table 7 shows the confusion matrix results, indicating that combining SMOTE and Tomek Links yields better-balanced performance than using a single method. In the negative class, out of a total of 672 data points, 665 were correctly predicted, only 1 was incorrectly predicted as neutral, and 6 were incorrectly predicted as positive. This indicates the model's excellent ability to recognize previously difficult-to-detect minority classes. For the neutral class, out of 670 data points, 490 were classified correctly, while 107 were incorrectly predicted as negative and 73 as positive. Although there are still quite a few errors, the number of correct data points is much better than the baseline or Tomek Links alone. In the positive class, out of 670 data points, 562 were correctly predicted, while 63 were incorrectly classified as negative and 45 as neutral. This performance is marginally worse than the baseline in the majority class; yet, it adheres to the balancing objective, resulting in equitable outcomes across all classes. The integration of SMOTE and Tomek Links mitigates bias towards the positive class and improves classification performance for the minority class, leading to more equitable performance across emotion categories.

Table 7. Confusion matrix using Naïve Bayes + SMOTE-Tomek Links

Class	Prediction results			Total
	Negative	Neutral	Positive	
Negative	665	1	6	672
Neutral	107	490	73	670
Positive	63	45	562	670

Table 8 indicates that integrating SMOTE and Tomek Links yields more consistent, balanced performance than the baseline and individual techniques. The model attained 79.64% precision, 98.96% recall, and an F1-score of 88.25% in the negative class. This indicates that nearly all negative data were accurately identified, albeit with a few classification inaccuracies. The neutral class achieved a precision of 91.42%, a recall of 73.13%, and an F1-score of 81.25%. Despite neutral recollection remaining comparatively inferior to other categories, this outcome significantly surpasses the baseline and the Tomek Links approach, which inadequately identified this class. The precision for the positive class was 87.67%, the recall was 83.88%, and the F1-score was 85.73%, marginally below the baseline; however, aligned with the balancing purpose, ensuring that performance is not predominantly focused on the majority class.

Overall, the model's accuracy was 85.33%, with a macro-average F1-score of 85.07% and a weighted-average F1-score of 85.08%. The minimal difference between the macro and weighted values indicates that the model can distribute performance fairly across all classes. This aligns with previous confusion matrix results, where the negative class, which was previously difficult to recognize, is now predicted very well. In contrast, the positive class continues to perform well even though it no longer dominates.

Compared with the baseline results, SMOTE, and Tomek Links alone, the combination of SMOTE + Tomek Links proved to be the most effective. This approach equilibrates class distributions while reducing the amount of confusing data. The confusion matrix indicates that the model can more effectively identify minority classes while maintaining performance on majority classes. This combination attains a balance between precision and recall across all courses, with a macro-average value nearly equivalent to the weighted average. Thus, Naïve Bayes + SMOTE + Tomek Links produces the most stable, fair, and representative sentiment classification performance compared to all previous scenarios.

Table 8. Results of confusion matrix using Naïve Bayes + SMOTE-Tomek Links

Class	Precision	Recall	F1-score
Negative	79,64%	98,96%	88,25%
Neutral	91,42%	73,13%	81,25%
Positive	87,67 %	83,88%	85,73%
Accuracy			85,33%
Macro avg	86,24 %	85,32%	85,07%
Weighted avg	85,82%	85,33%	85,08%

Based on the experimental results in Table 9, the Naïve Bayes model (baseline) achieved an accuracy of 83.52%, with notably uneven performance across classes, particularly in the negative class, which yielded a recall of 28.95% and an F1-score of 36.06%. The application of SMOTE successfully increased recall for minority classes, especially negative (97.77%), and improved the macro-average F1-score to a more balanced level (82.81%), even though total accuracy decreased slightly to 83.10%. In contrast, the exclusive use of Tomek Links proved ineffectual, evidenced by an accuracy of merely 74.00% and a significantly low F1-score for the minority class (Negative 5%, Neutral 32%), as this method solely diminishes borderline data without enhancing the representation of the minority class. The integration of SMOTE and Tomek Links yields optimal results, achieving an accuracy of 85.33%, a high F1-score across all classes (Negative 88.25%, Neutral 81.25%, Positive 85.73%), and a uniform distribution of performance among classes, evidenced by the macro-average and weighted-average values being nearly identical.

The experimental results show notable differences in performance across resampling strategies. The results indicate that SMOTE substantially improves minority-class recall by increasing representation during training. Nevertheless, this enhancement results in a marginal decline in the majority-class precision, as synthetic samples may create class overlap. Tomek Links, in isolation, does not resolve this issue as it merely eliminates borderline samples without addressing the deficiency of minority instances. The hybrid SMOTE–Tomek method adeptly reconciles these trade-offs by integrating data augmentation with noise mitigation, thereby improving performance stability across classes.

Table 9. Comparison of results

Model	Class	Precision	Recall	F1-Score	Accuracy
Naïve Bayes	Negative	47,83%	28,95%	36,06%	83,52%
	Neutral	79,44%	76,58%	75,31%	
	Positive	87,06%	91,98%	89,45%	
Naïve Bayes + SMOTE	Negative	78,03%	97,77%	86,80%	83,10%
	Neutral	88,70%	71,17%	78,97%	
	Positive	85,06%	80,39%	82,66%	
Naïve Bayes + Tomek Links	Negative	100%	3,00%	5,00%	74,00%
	Neutral	97,00%	19,00%	32,00%	
	Positive	73,00%	100%	84,00%	
Naïve Bayes + SMOTE + Tomek Links	Negative	79,64%	98,96%	88,25%	85,33%
	Neutral	91,42%	73,13%	81,25%	
	Positive	87,67%	83,88%	85,73%	

4.5. Statistical validation

Statistical validation was performed using confidence intervals to assess the robustness of the experimental results. The Wilson score interval was used to find the 95% confidence interval (CI) for classification accuracy. This is a good way to derive proportion-based metrics from confusion matrices, especially when the data is unbalanced.

The baseline Naïve Bayes model, evaluated on 971 cases, attained an accuracy of 83.52% (95% CI: [81.06%, 85.72%]). The Naïve Bayes model, when integrated with SMOTE–Tomek Links, achieved an accuracy of 85.33% on 2,012 occurrences, accompanied by a tighter 95% confidence interval of [83.73%, 86.82%]. This indicates that it was more solid and dependable.

McNemar's test was used to see if the difference in performance between the models was statistically significant. However, because resampling techniques change the size of the effective evaluation set, predictions from different models were not paired at the instance level, which goes against the assumptions of McNemar's test. Consequently, this test was not directly utilized in this study. In the future, all models will be tested on the same held-out test set to enable paired statistical significance testing. The confidence interval analysis indicates that the proposed SMOTE–Tomek Links methodology offers a more stable and reliable improvement over the baseline Naïve Bayes model for imbalanced sentiment classification.

4.5. Discussion

This study's results align with Putra's findings, which showed that using SMOTE can help identify minority sentiment classes but reduces overall accuracy by a small amount [6]. Arsanto also confirmed that the SMOTE–Tomek Links combination works well for handling imbalanced data, especially for detecting credit card fraud. This aligns with the improvements observed in this study in sentiment analysis [21]. Suandi also said that hybrid resampling methods that use SMOTE, ADASYN, and Tomek Links can greatly improve classification accuracy, reaching up to 89%. This shows that hybrid methods perform better than single-resampling methods [22]. These studies collectively validate the efficacy of hybrid resampling in addressing class imbalance across various application domains.

The effectiveness of the SMOTE–Tomek Links configuration in this study is demonstrated by both traditional performance measurements and statistical validation. The 95% confidence interval analysis indicates that the Naïve Bayes + SMOTE–Tomek Links model is more accurate and stable than the baseline Naïve Bayes model, as evidenced by the narrower confidence interval. This indicates that the observed performance enhancement is consistent and not attributable to random variation, particularly in the classification of minority emotion categories.

Sentiment classification studies frequently use classifiers such as support vector machine (SVM), logistic regression (LR), and random forest (RF); however, these models are known to be susceptible to significant class imbalance and often require supplementary strategies to address it. Furthermore, ensemble-based techniques may incur increased computational overhead. Naïve Bayes, on the other hand, is more efficient and scalable. The results of this study show that hybrid resampling can effectively reduce its sensitivity to imbalanced data. The suggested Naïve Bayes + SMOTE–Tomek Links framework is a useful, computationally efficient approach for sentiment classification in unbalanced educational review datasets.

In practice, computational efficiency is just as important as classification performance. When combined with TF-IDF features, Naïve Bayes is easy to compute and can be used to make large-scale predictions. The extra work that SMOTE and Tomek Links add only happens during preprocessing and does not make inference more complicated. So, the proposed framework strikes a good balance between classification balance and computational efficiency, making it a good choice for large-scale sentiment analysis on educational platforms.

5. CONCLUSION

This study confirms that class imbalance substantially degrades Naïve Bayes performance in sentiment classification of Ruangguru application reviews, particularly in recognizing minority sentiment classes. Although the baseline model achieved an accuracy of 83.52%, its recall and F1-score for negative sentiments were considerably low, indicating strong bias toward the majority class.

Experimental results demonstrate that applying SMOTE improves minority-class recall but introduces a trade-off in majority-class performance, whereas Tomek Links alone is insufficient because it removes only borderline samples without compensating for the scarcity of minority data. In contrast, the hybrid SMOTE–Tomek Links approach provides the most balanced outcome, increasing accuracy to 85.33% and achieving closely aligned macro- and weighted-average F1 Scores (≈85%), indicating more equitable performance across sentiment classes.

Statistical validation using a 95% confidence interval further shows that the hybrid approach yields more stable performance than the baseline Naïve Bayes model. From a practical perspective, the proposed framework offers a computationally efficient, easily deployable solution for sentiment monitoring on digital education platforms, as resampling is applied only during preprocessing and does not increase inference complexity. Therefore, this approach is suitable for large-scale analysis of user feedback to support service evaluation and continuous improvement in educational applications, particularly under imbalanced sentiment conditions.

6. LIMITATIONS AND FUTURE WORK

This study has several limitations. The dataset is limited to reviews from a single educational application, which restricts cross-platform generalizability. The SMOTE–Tomek Links setup employs fixed parameters without sensitivity analysis, and synthetic samples may inadequately capture linguistic complexities such as sarcasm, implicit sentiment, or Indonesian–English code-mixing. Furthermore, the assessment was conducted offline and failed to account for real-time implementation challenges, such as concept drift. The use of synthetic data raises ethical concerns about the potential amplification of bias; hence, resampling should be viewed as an auxiliary preprocessing step rather than a substitute for authentic data. Future work will broaden the dataset to encompass various educational applications and examine

contextual and multilingual embeddings, including BERT and IndoBERT, to enhance the representation of semantic variance. The suggested resampling technique will be assessed using ensemble and alternative classifiers, such as SVM, logistic regression, random forests, and deep learning models, and will be investigated for incorporation into real-time sentiment monitoring systems for educational platforms.

ACKNOWLEDGMENTS

We emphasize that this article has not been published elsewhere and is not sponsored by any specific organization. No funding was received to conduct this research.

FUNDING INFORMATION

The authors declare no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Yo Ceng Giap	✓	✓			✓	✓		✓	✓	✓			✓	
Christoper Andrian			✓	✓		✓	✓	✓			✓	✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The author declares that there is no conflict of interest in the research conducted.

DATA AVAILABILITY

The dataset used in this study comes from public datasets and is available on the website https://www.kaggle.com.

REFERENCE

[1] B. Santhi and G. R. Brindha, "Multinomial Naïve Bayes using similarity based conditional probability," *Journal of Intelligent and Fuzzy Systems*, vol. 36, no. 2, pp. 1431–1441, 2019, doi: 10.3233/JIFS-181009.
[2] D. K. Dake and E. Gyimah, "Using sentiment analysis to evaluate qualitative students' responses," *Educ Inf Technol (Dordr)*, vol. 28, no. 4, pp. 4629–4647, Apr. 2023, doi: 10.1007/s10639-022-11349-1.
[3] H. S. N. F. Masoud AminiMotlagh, "A reliable sentiment analysis for classification of tweets in social networks," *Soc Netw Anal Min*, vol. 13, no. 7, Dec. 2022.
[4] M. Umer *et al.*, "Scientific papers citation analysis using textual features and SMOTE resampling techniques," *Pattern Recognit Lett*, vol. 150, pp. 250–257, Oct. 2021, doi: 10.1016/j.patrec.2021.07.009.
[5] A. F. Rozi, A. Wibowo, and B. Warsito, "Resampling technique for imbalanced class handling on educational dataset," *JUITA: Jurnal Informatika*, vol. 11, no. 1, pp. 77–85, 2023.
[6] P. P. Putra, M. K. Anam, A. S. Chan, A. Hadi, N. Hendri, and A. Masnur, "Optimizing sentiment analysis on imbalanced hotel review data using SMOTE and ensemble machine learning techniques," *Journal of Applied Data Sciences*, vol. 6, no. 2, pp. 936–951, May 2025, doi: 10.47738/jads.v6i2.618.
[7] Barak, "Improving requirements classification with SMOTE-Tomek preprocessing," Jan. 2025, [Online]. Available: http://arxiv.org/abs/2501.06491.
[8] Md. M. Hossain, Z. Awosaf, Md. S. H. Prottoy, A. S. M. Alvy, and Md. K. Morol, "Approaches for improving the performance of fake news detection in bangla: imbalance handling and model stacking," *Proceedings of International Conference on Fourth Industrial Revolution and Beyond 2021*, pp. 723–734, 2022, doi: 10.1007/978-981-19-2445-3_51.
[9] J. Li *et al.*, *Advances in Sentiment Analysis - Techniques, Applications, and Challenges*, in Artificial Intelligence, vol. 22. IntechOpen, 2024. doi: 10.5772/intechopen.111293.
[10] M. Kumar, L. Khan, and H. T. Chang, "Evolving techniques in sentiment analysis: a comprehensive review," 2025, *PeerJ Inc.* doi: 10.7717/PEERJ-CS.2592.
[11] D. Elreedy, A. F. Atiya, and F. Kamalov, "A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning," *Mach Learn*, vol. 113, no. 7, pp. 4903–4923, Jul. 2024, doi: 10.1007/s10994-022-06296-4.

[12] A. Fernández, S. García, F. Herrera, and N. V Chawla, "SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary," *Journal of Artificial Intelligence Research*, vol. 61, pp. 863–905, 2018.
[13] L. Yang, P. Li, R. Xue, X. Ma, X. Li, and Z. Wang, "Intelligent classification model for railway signal equipment fault based on SMOTE and ensemble learning," in *IOP Conference Series: Materials Science and Engineering*, Institute of Physics Publishing, Jul. 2018. doi: 10.1088/1757-899X/383/1/012042.
[14] S. F. Taskiran, B. Turkoglu, E. Kaya, and T. Asuroglu, "A comprehensive evaluation of oversampling techniques for enhancing text classification performance," *Sci Rep*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-05791-7.
[15] E. F. Swana, W. Doorsamy, and P. Bokoro, "Tomek link and SMOTE approaches for machine fault classification with an imbalanced dataset," *Sensors*, vol. 22, no. 9, May 2022, doi: 10.3390/s22093246.
[16] W. Chandra, B. Suprihatin, and Y. Resti, "Median-KNN regressor-SMOTE-Tomek Links for handling missing and imbalanced data in air quality prediction," *Symmetry (Basel)*, vol. 15, no. 4, Apr. 2023, doi: 10.3390/sym15040887.
[17] C. Zhang *et al.*, "An empirical study on the joint impact of feature selection and data resampling on imbalance classification," *Applied Intelligence*, vol. 53, no. 5, pp. 5449–5461, Mar. 2022, doi: 10.1007/s10489-022-03772-1.
[18] M. Carvalho, A. J. Pinho, and S. Brás, "Resampling approaches to handle class imbalance: a review from a data perspective," *J Big Data*, vol. 12, no. 1, Dec. 2025, doi: 10.1186/s40537-025-01119-4.
[19] D. Elreedy and A. F. Atiya, "A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance," *Inf Sci (N Y)*, vol. 505, pp. 32–64, Dec. 2019, doi: 10.1016/j.ins.2019.07.070.
[20] A. Glazkova, "A comparison of synthetic oversampling methods for multi-class text classification," Aug. 2020, [Online]. Available: http://arxiv.org/abs/2008.04636.
[21] A. T. Arsanto, A. Faizin, M. Lutfi, Z. N. Saadah, and U. Y. Pasuruan, "Optimization of the Naive Bayes algorithm with SMOTETomek combination for imbalance class fraud detection," 2024. [Online]. Available: http://sistemasi.ftik.unisi.ac.id
[22] F. Suandi *et al.*, "Enhancing sentiment analysis performance using SMOTE and majority voting in machine learning algorithms," 2024, pp. 126–138. doi: 10.2991/978-94-6463-620-8_10.
[23] C. Zhao, Z. Yan, X. Sun, and M. Wu, "Enhancing aspect category detection in imbalanced online reviews: An integrated approach using Select-SMOTE and LightGBM," *International Journal of Intelligent Networks*, vol. 5, pp. 364–372, Jan. 2024, doi: 10.1016/j.ijin.2024.10.002.
[24] A. H. Luthfi, A. Faqih, and G. Dwilestari, "Enhancing model accuracy in sentiment analysis of the by.U application using Naïve Bayes and SMOTE techniques," *Journal of Artificial Intelligence and Engineering Applications*, vol. 4, no. 2, pp. 2808–4519, 2025, [Online]. Available: https://ioinformatic.org/
[25] I. N. Switrayana, D. Ashadi, H. Hairani, and A. Aminuddin, "Sentiment analysis and topic modeling of kitabisa applications using support vector machine (SVM) and SMOTE-Tomek links methods," *International Journal of Engineering and Computer Science Applications (IJECSA)*, vol. 2, no. 2, pp. 81–91, Sep. 2023, doi: 10.30812/ijecsa.v2i2.3406.
[26] R. Ardianto, T. Rivanie, Y. Alkhalifi, F. S. Nugraha, and W. Gata, "sentiment analysis on e-sport for education curriculum using Naïve Bayes and SVM," *Jurnal Ilmu Komputer dan Informasi*, vol. 13, no. 2, pp. 109–122, Jul. 2020, doi: 10.21609/jiki.v13i2.885.
[27] S. Ghosh, A. Hazra, and A. Raj, "A comparative study of different classification techniques for sentiment analysis," in *Research Anthology on Implementing Sentiment Analysis Across Multiple Disciplines*, IGI Global, 2022, pp. 174–183. doi: 10.4018/978-1-6684-6303-1.ch009.
[28] M. Raees and S. Fazilat, "Lexicon-based sentiment analysis on text polarities with evaluation of classification models," Sep. 2024, [Online]. Available: http://arxiv.org/abs/2409.12840.

BIOGRAPHIES OF AUTHORS



Yo Ceng Giap he completed his Bachelor's degree (S1) in Computer Science at STMIK Buddhi in 2003 and his Master's degree (S2) in Computer Science at STMIK Eresha in 2009. Currently, the author is enrolled as an active doctoral student in the Computer Science program at Dian Nuswantoro University in Semarang. Since 2004, he has been a permanent lecturer at Buddhi Dharma University in Tangerang, Banten, in the Computer Science program. He can be contacted at email: cenggiap@ubd.ac.id.



Christopher Andrian is a 2025 Bachelor's Degree student in the Informatics Engineering Study Program, Faculty of Science and Technology, Buddhi Dharma University. He has research experience in Computer Science. He can be contacted at email: christoper11a@gmail.com.